Pattern Recognition Letters 94 (2017) 68–74

Contents lists available at ScienceDirect

Pattern Recognition Letters

journal homepage: www.elsevier.com/locate/patrec

A simple approach to multilingual polarity classification in Twitter

Eric S. Tellez[a,c], Sabino Miranda-Jiménez[a,c,*], Mario Graff[a,c], Daniela Moctezuma[a,b], Ranyart R. Suárez[d], Oscar S. Siordia[b]

[a] CONACyT Consejo Nacional de Ciencia y Tecnología, Dirección de Cátedras, Insurgentes Sur 1582, Crédito Constructor, 03940, Ciudad de México, México
[b] Centro de Investigación en Geografía y Geomática "Ing. Jorge L. Tamayo", A.C. Circuito Tecnopolo Norte 117, Tecnopolo Pocitos II, 20313, Aguascalientes, México
[c] INFOTEC Centro de Investigación e Innovación en Tecnologías de la Información y Comunicación, Circuito Tecnopolo Sur 112, Tecnopolo Pocitos II, 20313, Aguascalientes, México
[d] División de Estudios de Posgrado, Facultad de Ingeniería Eléctrica, Universidad Michoacana de San Nicolás de Hidalgo, Santiago Tapia 403, 58000, Morelia, México

ARTICLE INFO

Article history:
Received 24 August 2016
Available online 22 May 2017

Keywords:
Multilingual sentiment analysis
Error-robust text representations
Opinion mining

ABSTRACT

Recently, sentiment analysis has received a lot of attention due to the interest in mining opinions of social media users. Sentiment analysis consists in determining the polarity of a given text, i.e., its degree of positiveness or negativeness. Traditionally, Sentiment Analysis algorithms have been tailored to a specific language given the complexity of having a number of lexical variations and errors introduced by the people generating content. In this contribution, our aim is to provide a simple to implement and easy to use multilingual framework, that can serve as a baseline for sentiment analysis contests, and as a starting point to build new sentiment analysis systems. We compare our approach in eight different languages, three of them correspond to important international contests, namely, SemEval (English), TASS (Spanish), and SENTIPOLC (Italian). Within the competitions, our approach reaches from medium to high positions in the rankings; whereas in the remaining languages our approach outperforms the reported results.

1. Introduction

Sentiment analysis is a crucial task in opinion mining field where the goal is to extract opinions, emotions, or attitudes to different entities (person, objects, news, among others). Clearly, this task is of interest for all languages; however, there exists a significant gap between English state-of-the-art methods and other languages. As expected some researchers decide to test the straightforward approach which consists in translating the messages to English, and then, use a high performing English sentiment classifier (for instance, see [3] and [4]), instead of creating a sentiment classifier optimized for a given language. However, the advantages of a properly tuned sentiment classifier have been studied for different languages (see, for instance [1,2,18,25]).

This manuscript focuses on the particular case of multilingual sentiment analysis of short informal texts such as Twitter messages. Our aim is to provide an easy-to-use tool to create sentiment classifiers based on supervised learning (i.e., labeled dataset); where the classifier should be competitive to those sentiment classifiers carefully tuned to a particular language. Furthermore, our second contribution is to create a well-performing baseline to compare new sentiment classifiers in a broad range of languages or to bootstrap new sentiment analysis systems. Our approach is based on selecting, using a search algorithm, a suitable combination of text-transforming techniques commonly used in Information Retrieval and Natural Language Processing such as n-grams of words and q-grams of characters, among others. The goal is that the text transformations selected optimize some performance measure, and the techniques chosen are robust to typical writing errors.

In this context, we propose a robust multilingual sentiment analysis method, tested in eight different languages: Spanish, English, Italian, Arabic, German, Portuguese, Russian and Swedish. We compare the performance of our approach in three international contests: TASS'15, SemEval'15-16 and SENTIPOLC'14, for Spanish, English and Italian respectively; the remaining languages are compared directly with the results reported in the literature. The experimental results locate our approach in good positions for all considered competitions; and excellent results in the other five

* Corresponding author at: INFOTEC Centro de Investigación e Innovación en Tecnologías de la Información y Comunicación, Circuito Tecnopolo Sur No 112, Fracc. Tecnopolo Pocitos II, Aguascalientes 20313, México .
E-mail addresses: eric.tellez@infotec.mx (E.S. Tellez), sabino.miranda@infotec.mx, sabinomiranda@gmail.com (S. Miranda-Jiménez), mario.graff@infotec.mx (M. Graff), dmoctezuma@centrogeo.edu.mx (D. Moctezuma), ranyart@dep.fie.umich.mx (R.R. Suárez), osanchez@centrogeo.edu.mx (O.S. Siordia).

http://dx.doi.org/10.1016/j.patrec.2017.05.024
0167-8655/